Adequate Condensed Representations of Patterns*

Arnaud Soulet[1] and Bruno Crémilleux[2]

[1]LI, Université François Rabelais de Tours
3 place Jean Jaurès
F-41029 Blois France
arnaud.soulet@univ-tours.fr
[2]GREYC-CNRS, Université de Caen
Campus Côte de Nacre
F-14032 Caen Cédex France
bruno.cremilleux@info.unicaen.fr

Patterns are at the core of the discovery of a lot of knowledge from data but their uses are limited due to their huge number and their mining cost. During the last decade, many works addressed the concept of condensed representation w.r.t. frequency queries. Such representations are several orders of magnitude smaller than the size of the whole collections of patterns, and also enable us to regenerate the frequency information of any pattern. Equivalence classes, based on the Galois closure, are at the core of the pattern condensed representations. However, in real-world applications, interestingness of patterns is evaluated by various many other user-defined measures (e.g., confidence, lift, minimum). To the best of our knowledge, these measures have received very little attention. The Galois closure is appropriate to frequency based measures but unfortunately not to other measures.

This paper extends the concept of pattern condensed representations. We propose a framework for condensed representations w.r.t. a large set of new and various queries named *condensable functions*. These queries encompass not only the frequency (conjunctive, disjunctive or negative) and frequency-based measures, but also address many other interestingness measures (e.g., minimum) and constraints having no suitable property of monotonicity. Condensed representations are achieved thanks to new closure operators automatically derived from each condensable function to get *adequate condensed representations*. We propose a sound and correct generic algorithm MicMac to efficiently mine the adequate condensed representations. Experiments show the conciseness of the adequate condensed representations, especially in dense and/or correlated data. They also demonstrate the scalability of our algorithm for measures or constraints which are intractible with naive methods.

We think that generalizing closure-based condensed representations will offer new tools for higher KDD tasks (e.g., non-redundant rules w.r.t. any measures), similarly there are many uses stemming from the frequency.

* This is an extended abstract of an article published in the Data Mining and Knowledge Discovery journal [1].

W. Daelemans et al. (Eds.): ECML PKDD 2008, Part I, LNAI 5211, pp. 20–21, 2008.

References

1. Soulet, A., Crémilleux, B.: Adequate condensed representations. Data Mining and Knowledge Discovery 17(1), 94–110 (August 2008)